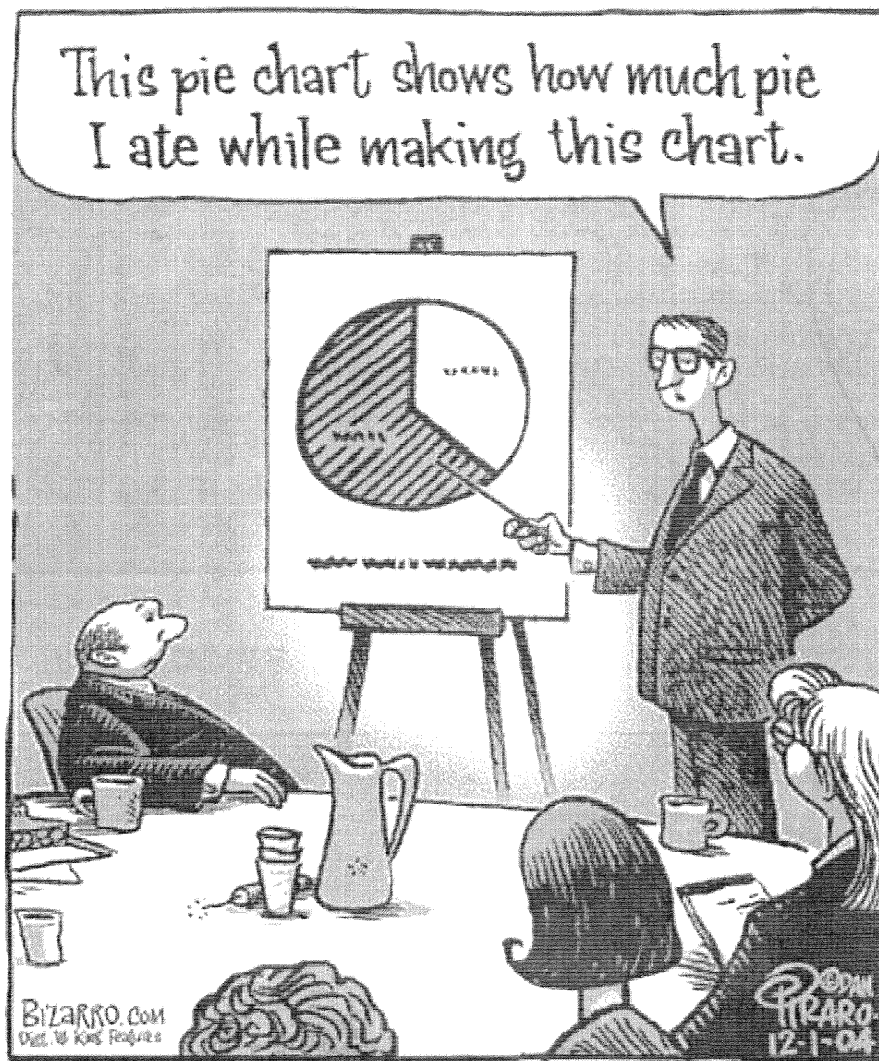


AP Statistics

Summer Assignment

Chapter 1: Exploring Data



Name _____

Introduction

statistics (p.4) –

~ any set of data contains information about some group of *individuals*, and this information is organized in *variables*

individuals (p.4) -

variables (p.4) -

- categorical variable (p.5) -

- quantitative variable (p.5)-

with a new set of data, we always ask:

- who (p.6) –

- what (p.6) –

- why (p.6) –

distribution of a variable (p.6) -

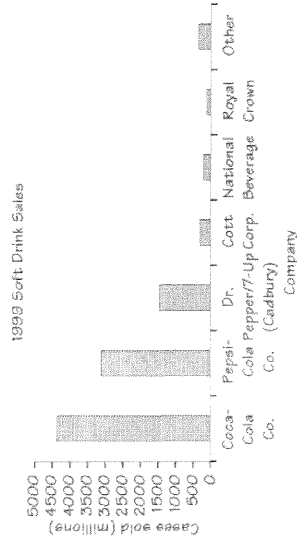
exploratory data analysis (EDA) (p.6) -

1.1 - displaying distributions with graphs

displaying categorical variables we use: bar graphs and pie charts

(p.8) distribution of a categorical variable lists the categories and gives either the _____ or the _____ of individuals who fall in each category.

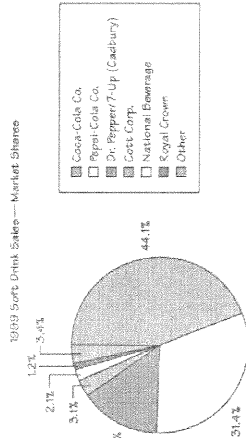
bar graph – quickly compares; heights of bars show the counts



(a)

- **remember:** label axis, title graph, scale axis, leave space between bars

pie chart - helps to see what part of the whole each group forms



(b)

- **remember:** must include all categories that make up a whole (= 100%)

advantages: helps an audience grasp the distribution quickly

disadvantages: takes time and space

1.6. In 1997 there were 92,353 deaths from accidents in the United States. Among these were 42,340 deaths from motor vehicle accidents, 11,858 from falls, 10,163 from poisoning, 4,051 from drowning, and 3,601 from fires.

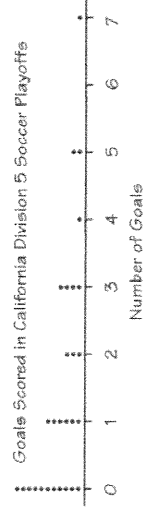
a. Find the percent of accidental deaths from each of these causes, rounded to the nearest percent. What percent of accidental deaths were due to other causes?

b. Make a well-labeled bar graph of the distribution of causes of accidental deaths. Be sure to include an "other causes" bar.

c. Would it also be correct to use a pie chart to display these data? If so, construct the pie chart. If not, explain why not.

* Note * The AP exam generally does not give you graph paper to construct graphs, get used to making them on blank paper like this one.

displaying quantitative variables we use: dotplots, stemplots, histograms
dotplots - one of simplest types of graphs to display quantitative data



• remember: label axis, title graph, scale axis

stem (and leaf) plot - graphs quantitative data

```

1 | 5 5 6
2 | 0 3 3 3 4 5 5 6 6 7 7 7 8 8 8 8 9 9
3 | 1 1 3 5 5 6 7 7 7 8
4 | 3 3 7
  
```

CAFFEINE CONTENT (MG) PER 8-OUNCE SERVING OF VARIOUS SOFT DRINKS

Key
3|5 means the soft drink contains 35 mg of caffeine per 8-ounce serving

• how to construct a stemplot:

1. separate each observation into a stem consisting of all but the rightmost digit (leaf)
2. write stems vertically in increasing order and draw a vertical line to their right, write each leaf to right of its stem
3. write stems again and rearrange leaves in increasing order out from stem
4. title graph and add key describing what stems and leaves represent

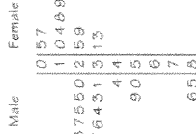
• split-stem and leaf plot - when splitting stems, be sure each stem is assigned an equal number of possible leaf digits

```

1 | 5 5 6
2 | 0 3 3 3 4 4
2 | 5 5 6 6 7 7 7 8 8 8 8 9 9
3 | 1 1 3
3 | 5 5 6 6 7 7 7 8
4 | 3 3
4 | 7 7
  
```

(b)

• **back-to-back stemplots** - good to compare two sets of data



• **remember**: 5 stems is a good minimum

advantages: easy to construct; displays actual data values

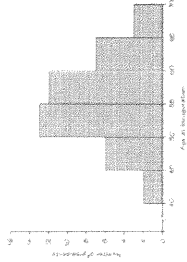
disadvantages: does not work well with large data sets

1.10 DRP TEST SCORES There are many ways to measure the reading ability of children. One frequently used test is the Degree of Reading Power (DRP). In a research of third-grade students, the DRP was administered to 44 students. Their scores were:

40 26 39 14 42 18 25 43 46 27 19
 47 19 25 35 34 15 44 40 38 31 46
 52 25 35 35 33 29 34 41 49 28 52
 47 35 48 22 33 41 51 27 14 54 45

Display these data graphically. Write a paragraph describing the distribution of DRP scores.

histograms - most common graph for distribution of one quantitative variable



• how to construct a histogram:

1. **divide range of data into classes of equal width and count number of observations in each class**; be sure to specify classes precisely so that each observation falls into exactly one class
2. **label and scale axis and title graph**
3. **draw a bar that represents the counts in each class**

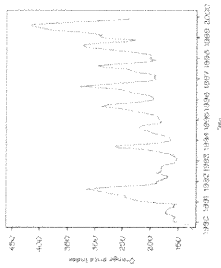
• **remember**: leave no space between bars; add a break-in-scale symbol (/) on an axis that does not start at 0; 5 classes is a good minimum

1.14 In 1993, *Forbes* magazine reported the age and salary of the chief executive officer (CEO) of each of the top 59 small businesses. Here are the salary data, rounded to the nearest thousand dollars:

118	121	162	208	362	424	539	736	291	58	498	643	360	322
180	368	664	244	896	300	343	536	545	217	295	1163	406	254
632	204	266	250	21	298	350	800	726	370	336	291	808	543
149	550	242	198	313	596	317	482	155	802	300	292	573	388
256	346	779											

Construct a histogram for this data.

time plots - plots each observation against the time at which it is measured



• trend (p.32) -

• seasonal variation (p.32)

• **remember:** mark time scale on horizontal axis and variable of interest on vertical axis; connecting points helps show pattern of change over time

When we describe a graph, we want to discuss the overall pattern of its distribution. Be sure to include the following terms

• center (p.12)

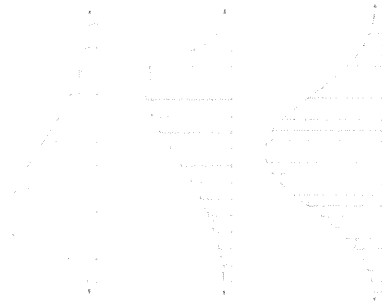
• spread (p.12)

• shape (p.12)

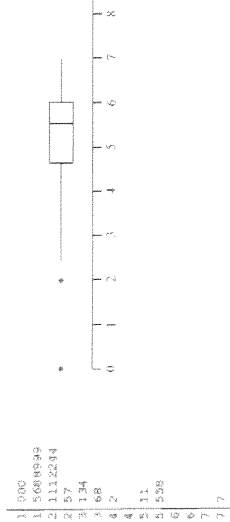
• skewed to the right (positively skewed) (p.25)

• skewed to the left (negatively skewed) (p.25)

• symmetric (p.25)



Outlier (p.12) -



percentile (pth percentile) (p.28) -

relative frequency (p.29) -

cumulative frequency (p.29) -

relative cumulative frequency (p.29) -

Age Group	Relative Frequency	Cumulative Frequency	Relative Cumulative Frequency
0-10	0.10	0.10	0.10
11-20	0.20	0.30	0.30
21-30	0.30	0.60	0.60
31-40	0.20	0.80	0.80
41-50	0.10	0.90	0.90
51-60	0.05	0.95	0.95
61-70	0.02	0.97	0.97
71-80	0.01	0.98	0.98
81-90	0.01	0.99	0.99
91-100	0.01	1.00	1.00

relative cumulative frequency graph/ogive - graph of relative standing of an individual observation

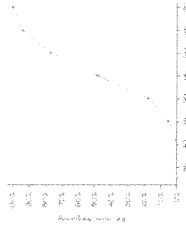


figure 1.12

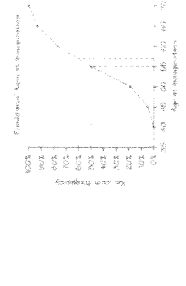


figure 1.13

• **remember:** begin ogive with a point at height 0% at left endpoint of lowest class interval, last point plotted should be at a point of 100%

1.29 Professor Moore, who lives a few miles outside a college town, records the time he takes to drive to the college each morning. Here are the times (in minutes) for 42 consecutive weekdays, with the dates in order along the rows:

8.25	7.83	8.30	8.42	8.50	8.67	8.17	9.00	9.00	8.17	7.92
9.00	8.50	9.00	7.75	7.92	8.00	8.08	8.42	8.75	8.08	9.75
8.33	7.83	7.92	8.58	7.83	8.42	7.75	7.42	6.75	7.42	8.56
8.67	10.17	8.75	8.55	8.67	9.17	9.08	8.83	8.67		

- Make a histogram of these drive times. Is the distribution roughly symmetric, clearly skewed, or neither? Are there any clear outliers?
- Construct an ogive for Professor Moore's drive times.
- Use your ogive from *b.* to estimate the center and 90th percentile for the distribution.
- Use your ogive to estimate the percentile corresponding to a drive time of 8.00 minutes.

1.2a - describing distributions with numbers

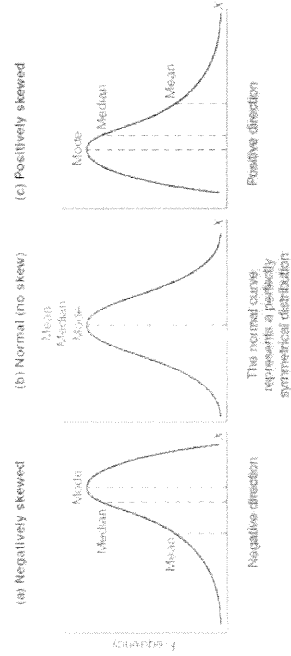
measuring center

• mean (\bar{x}) (read x-bar) (p.37) -

• median (M) (p.39) -

-
-
-

resistant measure (of center) (p.40) -



Examples of normal and skewed distributions

Example 1.32

The Survey of Study Habits and Attitudes (SSHA) is a psychological test that evaluates college students' motivation, study habits, and attitudes toward school. A private college gives the SSHA to a sample of 18 of its incoming first-year women students. Their scores are:

154 109 137 115 152 140 154 178 101
103 126 126 137 165 165 129 200 148

- Make a stemplot of these data. The overall shape of the distribution is irregular, as often happens when only a few observations are available. Are there any potential outliers?
- Find the mean score from the formula for the mean.
- Find the median of these scores. Which is larger: the median or the mean? Explain why.

measuring spread/variability

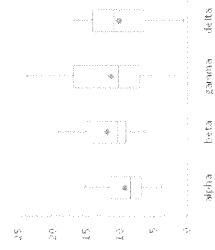
- range (p.42) -
- quartiles (p.42) -
 - first quartile (Q1) -
 - third quartile (Q3) -
 - "second quartile" -

interquartile range (IQR) (p.43) -

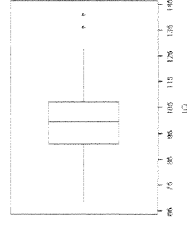
outliers (p.44) -

five number summary (p.44) -

box (and whiskers) plot - graph of five-number summary of a distribution; best for side by side comparisons since they show less detail than histograms; drawn either horizontally or vertically



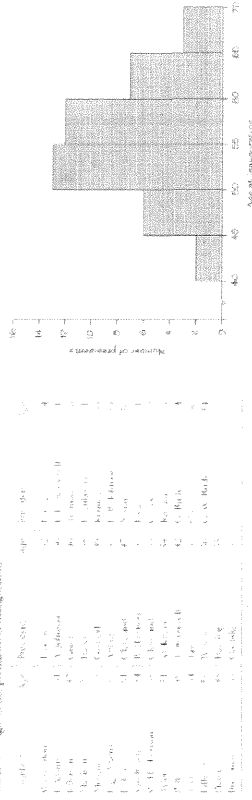
modified boxplot - plots outliers as isolated points; extend 'whiskers' out to largest and/or smallest data points that are not outliers



- **remember:** label axis, title graph, scale axis

1.37 Below you are given all the ages of the presidents at inauguration.

TABLE 1.4 Ages of the presidents at inauguration



a. From the shape of the histogram, do you expect the mean to be much less than the median, about the same as the median, or much greater than the median? Explain.

b. Find the five-number summary and verify your expectation from a.

c. What is the range of the middle half of the ages of new presidents?

d. Construct by hand a modified boxplot of the ages of new presidents.



e. On your calculator, define Plot1 to be a histogram using the list and Plot2 to be a modified boxplot also using the list. Graph. Is there an outlier? If so, who was it?

1.2b - describing distributions with numbers

measuring spread when we use the median as our measure of center

- range
- interquartile range

measuring spread when we use the mean as our measure of center

- standard deviation (p.49) –

the variance (s^2) of a set of observations is the average of the squares of the deviations of the observations from their mean

$$s^2 = \frac{1}{n-1} \sum (x_i - \bar{x})^2$$

the standard deviations (s) is the square root of the variance (s^2)

$$s = \sqrt{\frac{1}{n-1} \sum (x_i - \bar{x})^2}$$

degrees of freedom (p.50)–

Properties of standard deviations (p.50)

1.

2.

3.

1.41 New York Yankee Roger Maris held the single-season home run record from 1961 until 1998. Here are Maris's home run counts for his 10 years in the American League:

14	28	16	39	61	33	23	26	5	13
----	----	----	----	----	----	----	----	---	----

- Maris's mean number of home runs is $\bar{x} = 26.1$. Find the standard deviation s from its definition.
- Use your calculator to verify your results. Then use your calculator to find \bar{x} and s for the 9 observations that remain when you leave out any outlier(s). How does the "outlier" affect the values of \bar{x} and s ? Is s a resistant measure of spread?

linear transformations (p.53)

changes the original variable x into the new variable x_{new} given by an equation of the form:

$$x_{new} = a + bx$$

adding the constant a :

multiplying by the positive constant b :

\sim linear transformations do not change the shape of a distribution

effects of linear transformations (p.55)

- linear transformations:
- multiplying each observation by a positive number b :
- adding the same number a (either positive or negative) to each observation:

1.44 Maria measures the lengths of 5 cockroaches that she finds at school. Here are her results (in inches):

1.4	2.2	1.1	1.6	1.2
-----	-----	-----	-----	-----

- Find the mean and standard deviation of Maria's measurements.
- Maria's science teacher is furious to discover that she has measured the cockroach lengths in inches rather than centimeters (There are 2.54 cm in 1 inch) She gives Maria two minutes to report the mean and standard deviation of the 5 cockroaches in centimeters. Maria succeeded. Will you?
- Considering the 5 cockroaches that Maria found as a small sample from the population of all cockroaches at her school, what would you estimate as the average length of the population of cockroaches? How sure of the estimate are you?

1.54 The mean \bar{x} and standard deviation s measure center and spread but are not a complete description of a distribution. Data sets with different shapes can have the same mean and standard deviation. To demonstrate this fact, use your calculator to find \bar{x} and s for the following two small data sets. Then make a stem-plot of each and comment on the shape of each distribution.

Data A	9.14	8.14	8.74	8.77	9.26	8.10	6.13	8.10	9.13	7.26	4.74
Data B	6.58	5.76	7.71	8.84	8.47	7.04	5.25	5.56	7.91	6.89	12.50

Name: _____

AP Statistics
Chapter 1 Review Questions

1. Which of the following statements is true?
 - i. Two students working with the same set of data may come up with histograms that look different.
 - ii. Displaying outliers is less problematic when using histograms than when using stemplots.
 - iii. Histograms are more widely used than stemplots or dotplots because histograms display the values of individual observations.
 - (a) I only
 - (b) II only
 - (c) III only
 - (d) I and II
 - (e) II and III
2. Which of the following are true statements?
 - i. The range of the sample data set is never greater than the ranges of the population
 - ii. The interquartile range is half the distance between the first quartile and the third quartile.
 - iii. While the range is affected by outliers, the interquartile range is not.
 - (a) I only
 - (b) II only
 - (c) III only
 - (d) I and II
 - (e) I and III
3. On August 23, 1994, the 5-day total returns for equal investments in foreign stocks, U.S. stocks, gold, money market funds, and treasury bonds averages 0.086% (Business Week, September, 5, 1994, p. 97). If the respective returns for foreign stocks, U.S. stocks, gold, and money market funds were +0.70%, -0.11%, +1.34%, and +0.05%, what was the return for treasury bonds?
 - (a) 0.4132%
 - (b) 0.43%
 - (c) -1.55%
 - (d) -1.636%
 - (e) -1.77%
4. A 1995 poll by the Program for International Policy asked respondents what percentage of the U.S. budget they thought went to foreign aid. The mean response was 18% and the median was 15%. What do these responses indicate about the shape of the distribution of these responses?
 - (a) The distribution is skewed to the left.
 - (b) The distribution is skewed to the right.
 - (c) The distribution is symmetric around 16.5%.
 - (d) The distribution is uniform between 15% and 18%.
 - (e) Not enough information.

5. A teacher is teaching two AP Statistics classes. On the final exam, the 10 students in the first class averages 92 while the 25 students in the second class averaged only 83. If the teacher combines the classes, what will the average final exam score be?
 - (a) 87
 - (b) 87.5
 - (c) 88
 - (d) None of the above
 - (e) Not enough information
6. A teacher was recording grades for her class of 32 AP Statistics students. She accidentally recorded one score much too high (she put a "1" in front, so the score was 192 instead of 92, which was the top score). The corrected score was greater than any other grade in the class. Which of the following sample statistics remained the same after the correction was made?
 - a. Mean
 - b. Standard Deviation
 - c. Range
 - d. Variance
 - e. Interquartile Range
7. During the early part of the 1994 baseball season, many sports fans and baseball players notice that the number of home runs being hit seemed to be unusually large. Here are the data on the number of home runs hit by American and National League teams.

American League	35	40	43	49	51	54	57	58	58
	64	68	75	77					
National League	29	31	42	46	47	48	48	53	55
	55	55	63	63	67				

 - (a) Construct back to back stemplots to compare the data.
 - (b) Calculate numerical summaries of the number of home runs hit in the two leagues.
 - (c) Are there any outliers in the two sets? Justify your answer numerically.
 - (d) Write a few sentences comparing the distributions of home runs in the two leagues.

8. Create a data set with five numbers in which

(a) The mean, median and mode are all equal

(b) The mean is greater than the median

(c) The mean is less than the median

(d) The is higher than the median or mean

(e) All the numbers are distinct and the mean and median are both zero

9. Mail-order labs and 1 hour minilabs were compared with regard to price for developing and printing one 24-exposure roll of film. Prices included shipping and handling. Following is a computer output describing the results:

Mail-order			
Mean = 5.37	Standard Deviation = 1.92		Min = 3.51
Max = 8.00	N = 18	Median = 4.77	
Quartiles = 3.92, 6.45			
Mini-labs			
Mean = 10.11	Standard Deviation = 1.32		Min = 5.02
Max = 11.95	N = 15	Median = 10.08	
Quartiles = 8.97, 11.51			

Draw parallel boxplots displaying the distributions. Compare the distributions. Are there any outliers?